DOCUMENT RESUME

ED 194 562 TM 800 691

AUTHOR Douglass, James B.
TITLE A Process for Testing a Mathematical Model for the Solution of a Practical Problem: Application to Test Equating. LES Paper on Learning and Teaching. Paper #79.
PUB DATE Apr 79
NOTE 16p.: Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores: Error of Measurement: Latent Trait Theory: *Mathematical Models: Pretests Posttests: *Problem Solving: Sampling: Scaling: Test Items: *Test Reliability: *Test Validity
IDENTIFIERS *Anchor Test Study: Classical Test Theory: Estimation: One Parameter Model: *Test Equivalence: Three Parameter Model

ABSTRACT

A general process for testing the feasibility of applying alternative mathematical or statistical models to the solution of a practical problem is presented and flowcharted. The system is used to develop a plan to compare models for test equating. The five alternative models to be considered for equating are: (1) anchor test equating using classical test theory; (2) anchor test equating using the one-parameter logistic model; (3) anchor test equating using the three-parameter logistic model; (4) preequating using the one-parameter logistic model: and (5) preequating using the three-parameter logistic model. Sources of estimation instability are identified for the logistic models. Various practical and theoretical problems associated with the general process and test equating are discussed. The equating problem is of interest in its own right and serves to explicate the general process. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194562

AUG 1980
RECEIVED
ERIC/TM

PAPERS ON LEARNING AND TEACHING

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 800691

Learning and Evaluation Service
Michigan State University
East Lansing, Michigan

A Process for Testing a Mathematical Model for the Solution of a Practical Problem: Applications to Test Equating

James B. Douglass

Paper # 79

MSU IS AN AFFIRMATIVE
ACTION/EQUAL OPPORTUNITY
INSTITUTION

A PROCESS FOR TESTING A MATHEMATICAL MODEL FOR THE SOLUTION OF A PRACTICAL PROBLEM: APPLICATIONS TO TEST EQUATING

JAMES B. DOUGLASS
Learning and Evaluation Service
Michigan State University

A general process for testing the feasibility of applying alternative mathematical or statistical models to the solution of a practical problem is presented and flowcharted. The system is used to develop a plan to compare models for test equating. The five alternative models to be considered for equating are:

1) Anchor test equating using classical test theory.

2) Anchor test equating using the one-parameter logistic model.

3) Anchor test equating using the three-parameter logistic model.

4) Preequating using the one-parameter logistic model.

5) Preequating using the three-parameter logistic model.

Sources of estimation instability are identified for the logistic models. Various practical and theoretical problems associated with the general process and test equating are discussed. The equating problem is of interest in its own right and serves to explicate the general process.

Testing Alternative Models

The application of the general process requires that the following four conditions be met.

1) There is a practical problem of interest which can be solved by applying a mathematical or statistical model.

2) The model rests upon explicit assumptions.

3) At least one procedure exists for estimating the parameters of the model.

4) The estimation procedure requires sampling from one or more populations.

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 8-12, 1979. The research upon which this paper is based was supported by the Educational Testing Service. The author is particularly indebted to Dr. Gary Marco for conceiving of and guiding the research.

Figure 1 contains a flowchart of the general process for testing alternative models. The numbered steps below refer to the numbered boxes in the flowchart.

Step 1: Specify models. The first step is to specify each alternative model in a form that will allow it to be compared with competing models. Each model should be investigated to determine its appropriateness for each practical application being considered.

In order to make the test equating example more concrete, two assumptions will be made.

1) An operational testing system is in existence which is using some form of classical test theory and anchor test equating.

2) All items can be pretested in the operational system and item parameters can be estimated.

The first assumption is for convenience. The second assumption is necessary to accomplish preequating. It is also assumed that the reader is familiar with basic concepts of latent trait theory (e.g., Hambleton and Cook, 1977).

The mathematical models used in the example of test equating consist of two parts. The first part of each model is the equating model and the second part is the item or test model. The equating models under consideration are preequating and anchor test equating. The item or test models are classical test theory, the one-parameter logistic model and the three-parameter logistic model. The primary combined model under consideration is preequating with the three-parameter logistic model. Comparison models include anchor test equating with classical test theory, anchor test equating with the one-parameter logistic model, anchor test equating with the three-parameter logistic model, and preequating with the one-parameter logistic model.

Anchor test equating requires that any two tests to be equated contain a common set of items. Equating can be accomplished only after both tests have been administered and scored. A test can be preequated only if all items on the test have been pretested. Item parameters can be estimated during pretesting and a test can be equated to previous tests before it is administered (Marco, 1977). Preequating is only possible if an item model is used that specifies a relationship between item parameters and test scores and if the item parameters are reasonably invariant over different groups of examinees.

There are actually several different models that could be considered to use anchor test equating and classical test theory (Anghoff, 1971). For the purposes of this example, it is assumed that one model has been chosen for comparison or that one is currently being used for equating. Traditional equating methods will not be discussed in any detail.

Both the one-parameter and three-parameter logistic item models assume that the probability of correct response to an item is a logistic function of ability. The one-parameter model assumes that item characteristic curves vary only in difficulty. On a graph with ability on the abscissa and probability of correct response to an

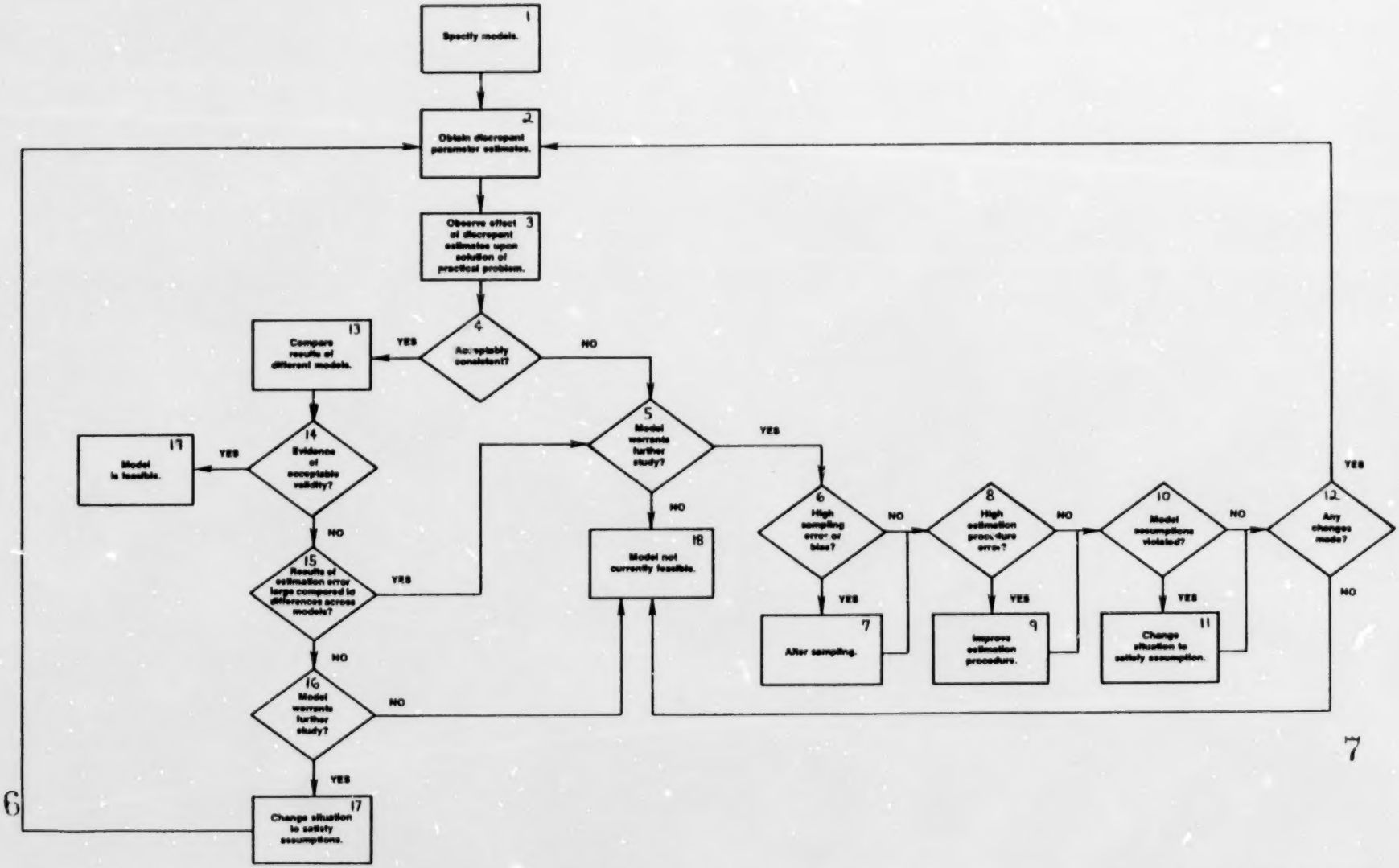

Figure 1: A process for testing the feasibility of a model for solving a practical problem.

item on the ordinate, item curves differ only by a translation on the abscissa. The lower asymptotes are zero and the slopes are equal at any given probability level. The one parameter associated with each item is the ability level at which the slope of the curve is greatest. This ability level is called the difficulty of the item.

The three-parameter logistic model has three parameters for each item. Item curves can differ in difficulty, lower asymptote and slope. The slope of each curve at its point of inflection is the discrimination of the item. The lower asymptote is the probability of an examinee of very low ability responding to the item correctly. For a more detailed discussion of the logistic models and graphical representations of item characteristic curves see Hambleton and Cook(1977).

To compare the one- and three-parameter models it is necessary to estimate the difficulty of each item using the one-parameter model and the difficulty, discrimination and lower asymptote using the three-parameter model. The LOGIST (Wood, Wingersky and Lord, 1976) computer program can be used to estimate parameters and examinee ability levels for the one-, two-, or three-parameter logistic models.

Step 2: Obtain discrepant parameter estimates. For each model estimates of the parameters of the model should be obtained assuming that the model is a true representation of reality. The estimates should be made under conditions similar to those that would actually occur in practice. Various sources of estimation error should be considered. Separate estimates should be made that will reflect the effects of the various sources of error that occur in practice. Identification of sources of error requires familiarity with the model being tested and the practical problems to which it may be applied.

In the equating example parameter estimates for the logistic models should be obtained under variations of conditions like those expected in the operational testing system. These variations may be thought of as possible sources of error in parameter estimation. Some of the sources of error that should be considered in a testing system include:

1) Sampling of examinees. Although the logistic models theoretically provide sample free parameter estimation, examinees should be chosen randomly from the sample of examinees at any given test administration. Sufficient numbers of examinees should be chosen to insure convergence of the LOGIST program and stability of parameter estimates. Various size examinee samples should be drawn and parameter estimates compared. Comparison samples could reasonable range from 200 to 5000 for both the one-parameter and the three-parameter models.

2) Sampling of items for scaling other items. Four aspects of item sampling should be considered.

   a) Number of items precalibrated compared to number of items to be placed on scale. Separate LOGIST runs will result in item scales that differ in origin and unit of measurement for the three-parameter model. For

a given run, LOGIST estimates abilities, disregards abilities greater than ±3 (or a user specified value) standard deviations from the mean and linearly transforms the mean of the remaining abilities to zero and the standard deviation to one. If items with known parameters are administered along with items with unknown parameters, LOGIST can place the latter items on the same scale as the former. It is only necessary to hold the difficulty parameters for the precalibrated items fixed at their previously estimated values.

The number of precalibrated items compared to new items in a given test administration can be critical. For example, the Test of English as a Foreign Language program is using the larger of 20 items or 50 percent of a test section for item scaling. In general at least 50 percent of the items should be precalibrated. Various numbers and percentages of precalibrated items should be drawn to compare estimation results.

b) Distributions of item difficulties among the precalibrated items. What type of distribution results in the most stable item parameter estimation and how great a difference does it make? In particular the distributions of items contained in current equating sections should be compared to other distributions.

c) Distributions of item types. Item types differ in efficiency and in ability measured. What is the effect on parameter estimation? In particular, items that violate assumptions of the model should be investigated. For example, items based on a common reading section probably violate the assumption of local independence.

d) Form to form variation. The effect of different test forms where numbers of items and distributions of item difficulties and types are held constant should be investigated.

3) Speededness of test sections and the effect of patterns of omits and no responses. Items with low numbers of examinees answering them may have a significant impact on parameter estimation. The pattern and frequency of omits and no responses will depend upon the speededness of the sections. Item calibration sections could either be shorter or items with high numbers of omits or no responses could be dropped. Further study of the effects of these factors upon estimation is needed.

Step 3: Observe effect of discrepant estimates upon solution of practical problem. The magnitudes of the differences in parameter estimates are important to the extent that they affect the solution of the practical problem (cf. Lord and Novick, 1968, p. 383). The effects of estimation discrepancies can be investigated in two ways. One method of determining estimation effects is to use the different parameter estimates obtained in the previous step to solve the practical problem. The resulting solutions would then be compared for consistency. If, for example, three independent sets of parameter estimates had been obtained, three practical solutions would be compared for consistency.

The second method of determining the effects of estimation is to identify a range of values for the various model parameters. The ranges would be determined from the discrepancies found in the previous step. A computer simulation could be

done by varying parameter estimates within the identified ranges. The resulting changes in the solution to the practical problem would then be investigated.

Variation associated with the sources of error discussed in the previous step will result in discrepant item parameter estimates for test equating. The first method of investigating the effects of the discrepancies is to use each set of item parameter estimates to independently equate a test. The resulting equatings can be compared for consistency. The second method is to use the discrepant estimates to determine variations that might be expected for each parameter for each item. A computer simulation of equating could be done with systematic variation in the item parameter estimates. The various equatings would be compared for consistency.

Step 4: Acceptably consistent? If it is determined that the variations in parameter estimation result in sufficiently consistent practical problem solutions, the validity of the results should be considered. If the solutions are not acceptably consistent or if ways of increasing consistency are desired and it is decided that the model warrants further study (Step 5), three possible sources of unstable estimates should be investigated. The sources of estimation instability are sampling error or bias, estimation procedure error, and violations of assumptions.

Step 6: High sampling error or bias? One cause of unstable parameter estimation is sampling error. The effect of increasing the sample sizes should be determined either theoretically or empirically. Of course it is desirable to have some knowledge of the effects of sample sizes before the initial parameter estimation so that reasonable sample sizes will be obtained initially. Sampling bias can also cause unstable estimates. Bias can occur as a result of choosing persons or conditions upon which estimations are based (see the discussion in Step 2 above).

In the equating example, if the equatings are not acceptably consistent and it is decided that further investigation of the model is warranted, sources of unstable estimates should be investigated. The first source of unstable estimation is the sampling error or bias. It may be possible to restrict the sources of variation in the operational system in order to reduce estimation discrepancies. For example, if the number of precalibrated items critically affects the stability of estimation, providing a sufficient number of precalibrated items at each administration will help stabilize the estimates.

Step 8: High estimation procedure error? Another source of unstable parameter estimates is the estimation procedure itself. Many mathematical models require computer programs using approximation procedures for parameter estimation. The procedures may result in estimations that fail to converge or converge to different results in different situations. Computer programs may place practical restrictions on estimates in order to increase the efficiency of estimation or the probability of convergence. These restrictions may result in biased or unstable estimates.

In the LOGIST program, for example, the lower asymptotes of the item curves of the three-parameter model are in general poorly determined. The program must often rather arbitrarily set these values. Fortunately the effect of the lower asymptote estimates is minimal except at very low ability levels.

Typically improvement of the estimation procedures can only be accomplished by experts in computer programming and the relevant mathematics. Estimation procedures

can often be improved based on the experience of using a model in an operational system. For example, the estimation procedures in the LOGIST program have been improved as a result of the TOEFL testing system using the three-parameter logistic model for preequating.

Step 10: Model assumptions violated? A third source of unstable parameter estimates is lack of fit of the model to the practical situation. Particular sources of lack of fit can be investigated by testing the various assumptions of the model in the practical situation. If any assumptions are greatly violated, it may be possible to alter the situation to meet the assumptions.

In the equating example both logistic models assume a unidimensional latent space. Unidimensionality implies that only one ability is being measured by the items in a test. Although no completely acceptable method for determining unidimensionality is available, a single factor appearing in the tetrachoric item correlations is a sufficient but not necessary condition (Lord and Novick, 1968, pp. 381-382). Although the assumption of unidimensionality is never entirely met in practice, it may be possible to obtain a more unidimensional test by making the items more homogeneous or by analyzing different sections of a test separately.

Local independence requires that the distributions of item scores for examinees of fixed ability be independent of one another. Local independence is equivalent to having specified a complete latent space (in this case unidimensional). The assumption of local independence will probably be violated by items based on a common reading passage or a common graph or figure. Removal of such items should help meet the assumption of local independence.

The one-parameter model assumes equal item discriminations. Outliers may be removed from the item pool in order to meet this assumption. More study of the assumptions of the models and their robustness needs to be done before the effect of violations of assumptions can be determined. In a practical situation it can be hoped that attempts to further meet assumptions will reduce estimation instability and increase the consistency and validity of the model.

Step 12: Any changes made? If no changes were made to improve the consistency of the results, the model is not currently feasible. Of course, improvements made in the future may make the model feasible. If measures were taken to improve stability, parameters should be reestimated with new samples. This is a return to Step 2. The effects of the new estimation discrepancies upon the practical problem should be investigated by simulation or by comparing practical solutions based on the discrepant parameter estimates. The solutions should be checked for acceptability. If the results are unacceptable further attempts may be made to reduce sources of estimation instability or the model may be deemed currently not feasible for the particular application. If the solutions are acceptably similar, the validity of the model should be investigated.

Step 13: Compare results of different models. Tests of the validity of the model depend on what types of comparisons are available. In the ideal situation the results of the model can be compared to an observable criterion under the various conditions likely to occur in practice. Unfortunately, observable criteria often are not available.

If no adequate observable criterion exists, the results of one model can be compared to the results of other models. At least six outcomes can result from model comparisons.

1) The results of the models converge in every situation of interest. This provides some indirect evidence of validity. Perhaps more importantly it implies that the decision of which model to use depends upon factors other than validity. Further discussion of this point follows under Step 19.

2) The results of the model being tested converge with those of another model or models where the latter show evidence of validity. This would occur in situations where the assumptions of the latter models are met and experience indicates consistently good results. If the model being tested makes less restrictive assumptions in the practical situation, then it can be argued that it will also work well in situations where its assumptions are met but it cannot be directly tested.

3) The results of the model being tested diverge from those of another model or models where the latter show evidence of validity. The validity of the model being tested is questionable in other situations as well.

4) The results of the model being tested converge with those of another model where the latter provides poor results. The validity of the model being tested is likely to be low.

5) The results of the model being tested diverge from those of another model where the latter provides poor results. The model being tested may or may not be valid.

6) The results of the models diverge in all situations. External criteria are needed to determine where each model is appropriate.

In the equating example, the patterns of convergence and divergence of equating results based on the different models would provide information useful in determining validity. For example, various methods of latent trait equating often give results similar to traditional equating where the assumptions of the traditional models are met (see Lord, 1977, p. 130 and Marco, 1977, p. 158). The latent trait models make less restrictive assumptions about the relationships between the tests to be equated. It can be argued that the latent trait methods are also likely to provide good results where their assumptions are met but the assumptions of classical equating are not (outcome 2 above).

Step 15: Results of estimation error large compared to differences across models? If there is not evidence of acceptable validity, the results of the estimation error within models may be investigated. The solutions to the practical problem may vary greatly as a result of parameter estimation compared to the solutions across models. If the model warrants further study, the various sources of estimation error should again be investigated. On the flowchart this is a return to Step 6.

If the results of estimation errors are not large compared to differences across models, it must be determined if the model warrants further study. There must be good reasons for continuing since the model fails to show evidence of validity despite consistent results and some evidence of stable parameter estimation.

Step 17: Change situation to satisfy assumptions. If it is decided to study the model further, greater efforts must be made to satisfy the assumptions of the model. After the situation is modified, a new set of discrepant parameter estimates should be obtained (Step 2) and the model should be retested by again following the general process outlined in the flowchart.

Step 19: Model is feasible. If a model is found to be acceptably consistent and there is evidence of acceptable validity, the model is feasible for the particular application being investigated. However, the reliability and validity of the model is dependent on the application being considered. Both the effects of parameter instability on the practical problem and comparisons between models were specific to the proposed application. The same model applied to a new problem would have to be reevaluated. Of course any commonalities between applications could be exploited to reduce the amount of new evaluation needed.

The reliability and validity of a given model provide useful information but are not sufficient for deciding between alternative models. Other considerations include:

1) Conceptual simplicity. The simpler models are preferable to more complex models if other factors are equivalent. For example, if the one-parameter and three-parameter models give essentially equivalent results in all applications of interest, the one-parameter model is to be preferred.

2) Costs of implementation and operation. The relative costs of introducing a model and long-term operational costs should be compared for alternative models.

3) Operational advantages and disadvantages. Different models will have different advantages and disadvantages in an operational system. For example, preequating removes the necessity of equating after test administration. This allows earlier score reporting and can reduce the probability of equating errors caused by time deadlines.

4) Range of situations to which the model may be applied for a given practical problem. For example, the logistic models are superior to classical equating models with respect to test restrictions. The logistic models do not assume parallel tests whereas classical models do. Stability of parameter estimates across situations and the effect on equating for the various models is also relevant.

5) Other practical situations to which the model may be applied. For the logistic models, once item parameters are estimated the models can be applied to such problems as constructing traditional tests, redesigning traditional tests, mastery testing, and two-stage testing (Lord, 1977).

Additional Considerations for Test Equating

Some additional practical and theoretical problems to be considered include:

1) Item calibration. The scale upon which the three-parameter logistic model item parameters are based is only unique up to a linear transformation. It is therefore necessary to obtain a common scale for all items. LOGIST routinely puts all items on a common scale when the item parameters are estimated in one computer run. Items from separate LOGIST runs will generally not be on the same scale. The best way to obtain item parameters on a common scale is to administer items already on the desired scale to examinees along with the items to be scaled. In the LOGIST run the difficulty parameters of the precalibrated items are held fixed to insure that the other items will be on the same scale. For each operational system a plan will have to be developed for item scaling.

2) Establishment of a system for recording and retrieving logistic item parameters.

3) Interim equating. If an equating procedure is currently being used in an operational testing program, a plan should be devised for gradual implementation of the alternative model. Schemes can be devised for retaining the traditional equating until the alternative model proves consistent and valid. A plan should be devised for testing the model as the system moves toward implementation. A well-designed plan should include item calibration and studies based on the flowcharted process presented above. The option of retaining traditional equating should remain until the alternative model is determined to be both feasible and desirable. Contingency plans should be devised based on possible results of the consistency and validity studies. The studies should be an integral part of the interim equating.

   Other practical applications of new models, such as test scoring, development and construction should be considered in the early planning stages. Feasibility studies for these applications should also be conducted during the interim equating.

4) Observed vs. true score equating for the logistic models (Lord, 1977, pp. 131-133). If true score equating is used for the three-parameter model, the problem of test scores below the sum of the lower asymptotes of the items must be addressed. Theoretically an examinee with little or no ability should receive a score that is simply the sum of the lower asymptotes of the items on the test. Of course, in practice an examinee may score below that sum.

5) Relationship between ability and current test scales. If logistic equating is being considered for a current testing program a relationship must be established between the scale currently in use and the ability scale of the logistic model (Marco, 1977).

6) Identification of items that do not fit the model. Procedures should be developed for identification and modification of items that do not fit whichever logistic model is being considered.

7) Quality control procedures. Systems to check the accuracy of item calibration, parameter estimate recording, and equating procedures should be developed. The purpose is to eliminate mistakes in the equating system.

Summary

The process presented and flowcharted can be used to test the consistency and validity of mathematical models for the solution of practical problems. The process is based on the assumption that the models have estimatable parameters. The critical features of the process include investigating the consistency and validity of results obtained from alternative models. Inconsistent results can occur because of sampling error or bias, estimation procedure error or violation of model assumptions. Validity can be investigated by comparison of model results to observable criteria or by consideration of the patterns of convergence and divergence of results of different models. A model is considered feasible if it shows evidence of consistency and validity.

References

Anghoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971, pp. 508-600.

Hambleton, R. K., and Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Lord, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Wood, R. L., and Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, N. J.: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N. J.: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.